Projeto de Formatura – 2023 – Press Release

PCS - Departamento de Engenharia de Computação e Sistemas Digitais

Engenharia Elétrica – Ênfase Computação

Tema: Aprimoramento de Question Answering por Texto em Linguagem Natural para SQL

No trabalho, investigamos a apresentação de conhecimento sintático e semântico a um modelo de linguagem, baseado na arquitetura transformer, durante o treinamento para a tarefa de tradução de linguagem natural para SQL. Modelos para essa tarefa, assim como alguns transformers aplicados a atribuições igualmente sensíveis à pergunta do usuário, isto é, nas quais é vital a correta assimilação da frase do usuário para geração da saída requerida, por vezes não interpretam corretamente a requisição (por algum ponto particular na frase) e, por isso, não geram a resposta pretendida pelo usuário. Os modelos não interpretam corretamente as relações entre os termos da frase e a natureza delas, por exemplo se pedirmos "Quais modelos de veículos foram produzidos no Brasil e pela montadora X", o modelo de tradução de linguagem natural para SQL em Português não interpreta a relação de adição da conjunção 'e', mas traduz na consulta SQL a forma para um OR.

Portanto, para que o modelo se aproxime da real intenção do usuário e, assim, gere a saída pretendida, é preciso que interprete corretamente as relações expressas pelos termos da frase e a natureza delas. Assim, apresentamos no treinamento, por uma infusão de conhecimento, as formas sintáticas e semânticas das frases também na entrada.

A forma sintática da frase refere-se às funções sintáticas desempenhadas por cada termo e as relações entre essas funções. Logo, a forma sintática apresenta o sujeito da frase, o verbo raiz, os objetos direto e indireto, entre outras funções. Para capturar essas informações, recorremos a um parser de dependências, que as gera na forma de uma árvore, a qual linearizamos e inserimos junto com a respectiva frase na entrada.

A representação semântica é representação apenas dos termos e relações que carregam o núcleo semântico da frase, o âmago da informação que ela transmite. Termos como determinantes, por exemplo, não são abrangidos, portanto (ao contrário da representação sintática). Para representarmos essa informação, recorremos à Abstract Meaning Representation (AMR). Também usamos um parser para gerar a correspondente árvore AMR da frase, a qual linearizamos e concatenamos à frase na entrada.

A tarefa alvo é a tradução de linguagem natural para SQL, por ser particularmente sensível à frase de entrada, como mostrado no exemplo sobre a montadora de carros. A interpretação das relações entre termos da frase e tipo delas é fundamental para que as traduzamos para as respectivas representações em SQL e, assim, tenhamos na saída a consulta que reflita a intenção da frase. Nessa tarefa, o benchmark é feito sobre a base de dados Spider, um grupo de perguntas com vários níveis de SQL para várias tabelas. Treinamos um modelo de controle, usando a base de dados Spider, com as perguntas traduzidas para Português, sem informação alguma, e dois modelos, um usando a base Spider traduzida para Português com informação sintática, e um usando a base Spider traduzida para Português com a informação semântica das frases.

As métricas de avaliação dos modelos foram exact-set-match, que compara o quão próxima é a consulta gerada da consulta de referência, e execution accuracy, que mede se o resultado de execução da consulta SQL é o mesmo que o da consulta de referência. Pela soma das duas, temos a pontuação. Pelo treinamento dos modelos, vimos resultados importantes para o uso da informação sintática e semântica.

Integrantes: Anton Bulle Labate

Professor(a) Orientador(a): Fabio G. Cozman